

5.10 EXTRACT DATA

Introduction

The Extract Data screen is an extension of the StEPS Browse Data screen, and is accessible within StEPS from either the Tools Main Menu or the Browse Data screen. It provides you with capabilities beyond that of the Browse Data screen, including the ability to save the results of your searches and extracts into permanent data sets and selection sets. These data sets and selection sets can then be used for processing in StEPS or for downloading to use in other applications. Like the Browse Data screen, this screen is read only, which means any changes you would need to make to the data would have to either be made interactively in the Review and Corrections screens.

Specifically, the StEPS Extract Data screen allows you to:

- Subset the data in a given file by setting a pre-condition in SAS code.
- Further subset data using a where clause.
- Obtain information on the structure (PROC CONTENTS) of the data set.
- Create new variables based on computations from existing variables.
- Obtain summary statistics on the data set and any subsets, such as number of observations meeting the defined criteria.
- Access SAS/INSIGHT to examine and manipulate the subset data.
- Save the results of subsetting operations to a permanent data set or selection set.

Accessing the Screen(s)

The Search/Extract screen may be reached in one of two ways. Both ways involve an intermediate step of identifying a data set from which to search/extract data. The first method is from the Tools main menu, using the “Extract” icon. Clicking on this icon will bring up the screen below, which is similar to the Browse Data screen, but whose function is to allow you to select a data set to take into the Search/Extract process.

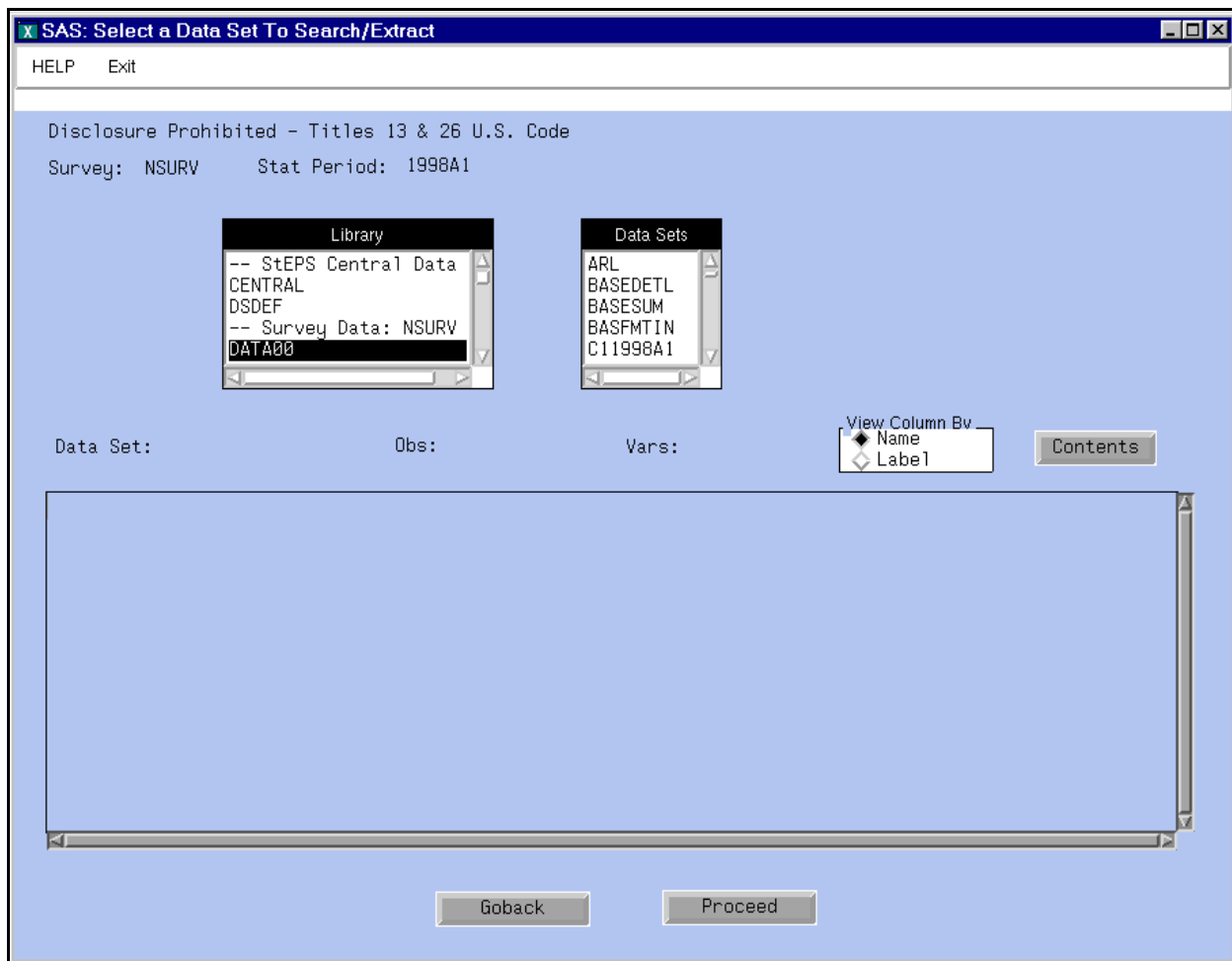


Figure 5.10.1 - Select a Data Set to Search/Extract screen

5.10.1 SELECTING A DATASET

- ! Select a data library from the “Library” window.
- ! Select a data set from the Data Sets window. The contents of the Data Set you selected will now appear in the data table window at the bottom of the screen.
- ! If you are ready to move on to the Extract screen, click on the “Proceed” button. If not, click on “Goback” to clear your selection and start over.

Example: You want to examine values from the 1998 stat period control file.

- ◆ Place the cursor from your mouse over the line “DATA00” in the “Library” window. Left click the mouse to select this line.
- ◆ Find the data set “C11998A1” in the “Data Sets” window. You may have to scroll to find this file. Place the cursor over this file and left-click the mouse to select it.

- ◆ The data set will now appear in tabular form in the data set portion of the window.
- ◆ Position the cursor over the “Proceed” button and left click the mouse.

The alternative way to enter the extract screen is through the Browse Data screen. You must already have a data set selected to browse before you can move into the Extract screen. (see Chapter 5.13 for guidance on entering and using the Browse Data screen).

! Click on the “Process” button in the Browse Data screen.

! Click on the “Access Search/Extract Screen” option from the pop-up menu.

Contents of the data set
Access Search/Extract Screen

! The following screen will appear:

Disclosure Prohibited - Titles 13 & 26 U.S. Code Survey: NSURV Stat Period: 1998A1 Date: 26FEB01:15:10:16

Source Data Set: DATA00.C11998A1 Obs: 6,811 Vars: 76 View Column By: Name Contents

	SURVEY	STATP	ID	PCFLG	SPLIT1	SPLIT2	SPLIT3	SPLIT4	SIC	SICRCD	NAICS	BRIDGE	SMPWG
1	NSURV	1998A	00510587455		519900	519999			350.0700
2	NSURV	1998A	00512231755		506590	506500			37.2145
3	NSURV	1998A	00512684855		506590	506500			6.5625
4	NSURV	1998A	00513265555		514130	514100			248.4000

Pre-Condition in SAS Code:

WHERE Condition in SAS Code:

Count Extract

Output Result Options

Selected: Tot 76 Available: Tot 76

Comp. Vars ACTION ANLREF BMFACT BMFCYC

Output: All Output limit to Sort By Reset

Output 1 in: 1

Result = Process Title: Result of Search/Extract Screen View Column By: Name Label

Figure 5.10.1 - Search/Extract Data by WHERE Clause Screen

Screen Features and Functionality

5.10.2 SPECIFYING EXTRACT CRITERIA

! The first step in the extraction process is to remove from the source data set those records (cases) you will not need in the extracted file. For example, if you are only interested in examining cases that are classified in NAICS 514130, the first step will be to remove all cases that are not classified within NAICS 514130. To subset the data set displayed in the “Source Data Set” window (see figure 5.10.1) use the “condition” windows immediately below it.

1. **Pre-Condition in SAS Code:** You can write a simple logic statement (in SAS syntax) to keep only those cases that meet the specified criteria. A precondition is not required, but may be used in conjunction with the required Where Condition to further subset your data set. To define a precondition, you may either set your equation to equal a value (or group of values) or to not equal a value (or group of values).

Example: NAICS = '514130' or NAICS in ('514130, '514131')
NAICS ≠ '514133' or NAICS not in ('514133','514134')

Note: Keep in mind that any variable (i.e. NAICS) used in the pre-condition must exist in the data set from which you are extracting records (the source data set). For example, you cannot subset the stat period (c1) control file using the variable NAME1, which exists only in the Master Control file (CONTROL).

2. **Where-Condition in SAS Code:** A Where Condition is required in order to take a data set through the extraction process. You may invoke the SAS Standard Where clause from the pick list of choices in order to define your condition, but you have other options as well. As in the Pre-Condition, any variable to be used in the Where clause must exist in the source data set.
 - a. There are multiple options available to set or define the where condition
 - i. Invoke the StEPS Standard Where clause
 - ii. Use criteria defined in edits or imputation: You may import the logic defined in one of your edit definitions into the where clause window.
 - iii. You have three options for retrieving a saved where clause:
 - (1) **saved in temporary memory from a previously run where clause:** You may recall to the where clause window a where clause you used earlier in the current session.
 - (2) **saved at survey level from previously run where clause or previous session:** You may call up a where clause you saved previously in another part of StEPS, such as a where clause used to define a selection set in the Review and Correction module.
 - (3) **saved at the user level (across surveys)**
 - iv. You may save a new where clause into either a temporary or permanent where

clause data set for future use.

- b. Example: WGT = 1.0000 or WGT < 20.0000
 - c. You MUST use a where clause in this screen
3. You may set multiple limiting conditions by using both the Pre-Condition and the Where Condition. Example:

Pre-condition: NAICS = '514130'

Where condition: ACTION ne 'D' and STATUS = 'A'

The cases extracted would be limited to just those meeting the three criteria specified above. In other words, the only cases you would keep in your extracted files would be those classified in NAICS 514130 that were active (STATUS = 'A') and not "deletes" (ACTION ne 'D')

Note: You can define multiple limiting conditions entirely within the where condition or using both the where condition and pre-condition. Sometimes, you may want to subset your file in several different ways to create different outputs. You will have to define a different where condition each time. For example, say you want to get all cases in NAICS 444130. Then you want to examine three different subsets of the data. First, you want to extract the active cases, then you want to extract the inactive cases, and finally you want to look at just the cases that have been ghosted. One advantage to using the pre-condition in combination with the where condition is that you could define the NAICS limitation in the pre-condition, then define each where condition (active, inactive, ghost) separately. Since the NAICS limitation is common to all three extracts, you only want to type this condition one time.

4. Additional Examples:

- a. You want to extract all cases from the 2000a1 stat period (C1) control file that are classified in NAICS 422310 and which have been checked in.

Pre-condition: NAICS = '422310'

Where condition: CKNDTE ne . (missing)

- b. You want to extract all cases from the 2000a1 stat period (C1) control file that are classified in NAICS 422311 that are not delinquent and which belong to panel 1.

Where condition: NAICS = '422311' and
PANEL = '1' and
RSPCDE = 'Y' (a response code of Y means the case is not delinquent)

- c. You want to extract only those records from the 2000 item file (it2000a1) for item C4PAY.

Where condition: ITEM = 'C4PAY'

- d. There may be situations where you want to keep all records in a file, but you are required to create a where condition in this screen. How do you work around this where clause requirement? You set your where condition such that every record/case in the data set will meet the criteria. For example, say you want to keep **all** cases from the 2000a1 stat period (C1) control file but are required by StEPS to use a where clause.

Where condition: STATP = '2000a1'



Note: This syntax will eliminate nothing from the data set you are extracting, since all records should have the value '2000a1' in the field STATP. However, we would recommend you always subset your data to make the extracted file smaller, especially if you intend to download the data set.

- ! After subsetting the source data set, the next step is to further limit the size of the output file. In the subset process, we limited the number of **records** to be included in the output data set. In this step, we are limiting the number of **variables** to be included on each record. For example, if we have subset our source data set to keep only ITEM C4PAY (see example c above), we may want to see only the reported data (RPDATA) for that item. Thus, we want to keep RPDATA and remove EDDATA, ADDATA, and WDDATA from our output data set. We may also want to remove additional variables beyond the data types, such as the data flags. All of this is done in the “Output Result Options” portion of the Search/Extract screen.

Note: You are not required to limit the number of variables you keep in your output data set. However, as stated previously, the smaller your output data set the easier it will be to manipulate and/or download. We recommend limiting your output unless you absolutely need every variable from the source data set.

1. There are four windows in the “Output Results Options” section.
 - a. **Selected:** This window displays those variables you have selected to keep in the output data set. The default value in this window is “_ALL_”, meaning that all variables found on the source data set will be included in the output data set. Unless you change this (see below) you will include all source variables in the output file.
 - b. **Available:** This window lists all the variables in the source data set individually. This is the list you will select from in deciding which variables to include in your output data set.
 - c. **Tot:** There are two “Tot” windows in the “Output Results Options” section. One is situated next to the Selected window and displays the total number of variables in this window. The second Tot window is situated next to the Available window and

displays the number of variables in this window. If the “_ALL_” option is displayed in the Selected window, the values in both Tot windows should be the same. For example, in Figure 5.10.1, you will see that both Tot windows show the value 76. This means there are 76 variables in the source data set and currently there are 76 variables selected for inclusion in the output data set.

2. You may select variables from the source data set for inclusion in the output data set. To do this, highlight a variable within the Available window and click the  between the two windows. This will transfer the variable to the Selected window, thus designating the variable for inclusion in the output data set. Repeat this process until you have placed all variables you wish to include in the Selected window. Note that ID is always included in the extracted file, so you do not need to specifically select this variable.
3. If you make a mistake and inadvertently select a variable (or variables) you do not wish to include, you have two options available to correct the problem:
 - a. You may click the “Reset” button, which is situated between the Select and Available windows. Doing this will clear **all** variable selections you have made to that point. You will then start over making your variable selections.
 - b. You may highlight a variable within the Selected window, then click the . This will deselect the highlighted variable. You may repeat this process as needed until you have the appropriate selection of variables in the Selected window.

! You may also generate “computed” variables for inclusion in your output file. You will create these variables by making them a “function” of existing variables in your source data set. For example, you can create a new variable that is the sum of EDDATA and WDDATA for each ITEM you extract. You may only create computed variables from numeric variables, thus this feature is pretty much limited to RPDATA, EDDATA, ADDATA, and WDDATA, dates, or various other numeric StEPS variables. You cannot produce a variable that is the sum of data from two different item records. Therefore, the primary utility of this feature would be to combine data values for a single item, such as the EDDATA and WDDATA values for C1PAY. You are not required to create computed variables; this is just another feature made available for you to “customize” your output data set.

1. Click on the “Comp. Vars” button, which is situated between the Selected and Available windows. The following screen will appear:

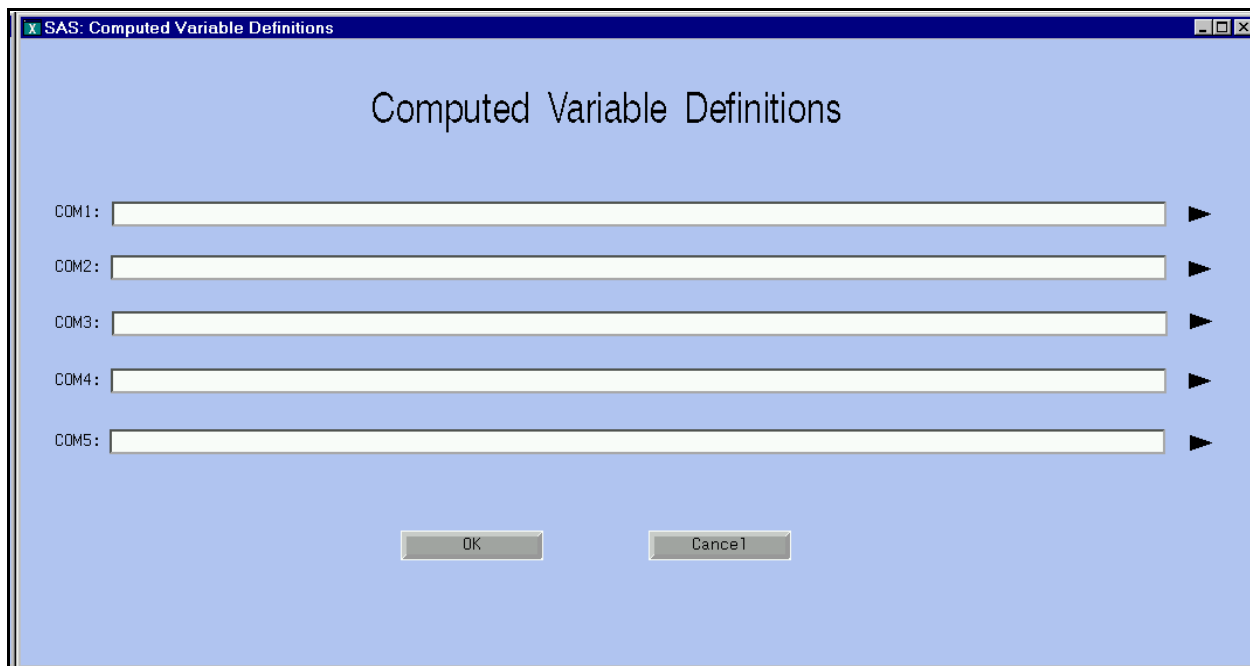


Figure 5.10.2 - Computed Variable Definitions screen

2. Key in the formula to create the computed variable:

Example: $COM1 = EDDATA + ADDATA;$

Note: The semicolon at the end of the formula is critical. Since you are essentially writing part of a SAS data step, the formula must comply with SAS syntax.

3. The picklist to the right of each computed variable window offers two options:
 - a. **Get Default Code to Start With:** Selecting this option produces an example of how the computed variable definition should appear. The example StEPS provides is $COM1 = COM1;$
 - b. **Clear Code:** This options allows you clear any formula you have entered and start over.
4. When you are satisfied with your computed variable definitions, click on the "OK" button. If you want to scrap your definitions and forget the whole thing, click the "Cancel" button. You will not see the results of your computed variable until you click the "Extract" button.
5. Here is an example of including computed variables in your extracted data set. Let's say you want to look at the sum of EDDATA and ADDATA for item variable C1PAY.

Precondition: None

Where Condition: ITEM = 'C1PAY'

Computed Variable: COM1 = EDDATA + ADDATA;

! At the start of the extract process, you were required to subset your source data set using at least a where clause and possibly a pre-condition. When you subset in this fashion, you are removing certain cases based on the data values within the source data set. In the output window, you have the option to limit the output even further. This subset, however, is based on some criteria you specify that has nothing to do with the data values themselves. For example, you may tell StEPS that you want to extract only the first 100 records that meet the where and pre-conditions. Or, you can tell StEPS to extract one record out of every 10 that meet the where conditions. As with the pre-condition and computed variables, you are not required to limit your output.

1. This feature is nice if you want to look at the output in the display window before deciding to save the data set permanently. If your output data set will be particularly large (i.e. >10,000 records) you may want to look at the first 100 or so records before deciding to proceed with download or SAS Insight.
2. The first step in limiting output is to select the “Output limit to” option in the “Output” portion of the screen. Note that the default value is “All”. If you select this option, every record that meets the where and pre-conditions will be included in the extracted data set. In other words, you will have done nothing to reduce the number of records in your output beyond what you have already specified in the where and pre conditions. If, however, you click on the “Output limit to” option, a field/window will appear in which you may designate the number of records to retain in the output:



- a. You may key a numeric value in this field. The number you enter will be the number of records that appear in the output data set. For example, if you key in 100, only the first 100 records meeting the where/pre conditions will be included in the output.
 - b. Alternately, you may click on the ➤ next to this field to bring up a pick list of values. These values include 1, 10, 100 and 500, then increase to 5000 in increments of 500 (i.e. 1000, 1500, 2000, etc.). You are not limited to the values in this pick list (you may key in any value you wish). The pick list just gives you a ready reference of commonly used values.
3. Another way to limit your output would be to include “every other” record or one out of every ten records. There is a field/window labeled “Output 1 in: “ at the bottom of the Output portion of the screen, and if you key a value in this field (say 100, for example), then StEPS will keep one record out of every 100 records which meet the where/pre conditions for inclusion in the output. There is a pick list of options for this field, running from 1 to 100. As with the “Output limit to” field, you are not limited to the values included in this field, as you may key in any value you want.

4. As a reminder, you are not required to limit your output in either of the two ways described in this section, but it may make reviewing the data easier before deciding if you wish to make your extract permanent.
- ! The final way you may modify your output is to define a sort order for the extracted data. You will find two fields within the “Sort” box which allow you define a sort order. The field to the left will have a white background, indicating it is correctable. The field to the right will have a gray background, indicating it is not correctable. The second field is dependent on the value in the first field, thus the second field will remain non-correctable until you enter a value in the first field. To define a sort order:
1. Enter the name of a variable from the source data set into the first field under “Sort”. This variable name must be one you designated to be included in the output. In other words, if you selected only ID, ACTION, and STATUS for inclusion in your output data set, you cannot sort based on NAICS! To select the name of a sort variable:
 - a. Key the variable name directly in the field, or
 - b. Use the pick list under the ► to the right of the field. This pick list will be limited to the variables you chose to include in your output under “Output Result Options”
 2. Select the sort order using the second field. You have only two options: Ascending and Descending. You may use the pick list to select the option you want and this will probably be easier than keying the value in yourself.
- ! The “Reset” button clears all sort and limit output options.

5.10.3 EXTRACTING THE DATA

- ! Now that you have finished defining all of your limiting conditions, you are ready to actually extract and create the output data set. You will do this by clicking on the “Extract” button which is situated to the right of the “WHERE Condition in SAS Code” window. The extracted data set will now appear in the data table display at the bottom of the screen.
1. You can determine how many records are contained in the out put data set by checking the value in the “Result =” field. You may also determine this by clicking on the “Count” button.
 2. You may give your output data set a title by keying this information into the “Title:” field. As a default, StEPS will call your new file the title “Result of Search/Extract Screen”.
 3. You may use the “View Column By” box on the right-hand side to adjust the display table. This box gives you two options: You may view the table columns by the variable names (this is the default setting) or by the “label” associated with the variable, as defined in the SAS data set. Selecting the “Label” option changes the display only. It does not change the names of the variables in the output data set.

5.10.4 OTHER OPTIONS

- ! You can examine the contents and structure of your source data set using the “Contents” feature. Click on the “Contents” button in the upper right-hand corner of the Extract screen to bring up the following screen:

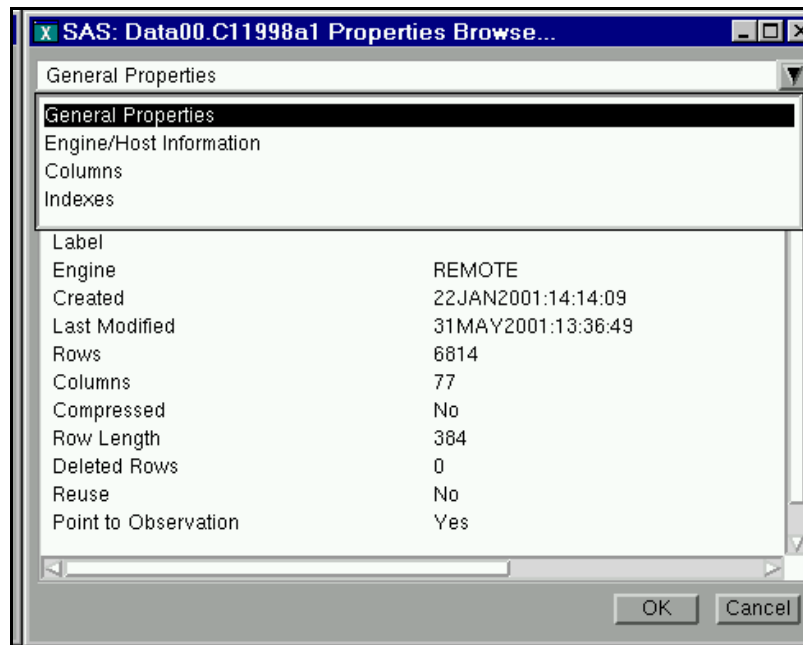


Figure 5.10.3 Properties Browse screen

1. Like a PROC CONTENTS in SAS, this screen provides structural information about the file itself. You may find this information useful in deciding how best to subset the file and which variables to keep. The specific information provided includes:
 - a. The NAME of each variable in the data set. By using the “VIEW” pmenu, you can change the order in which the variable names are displayed.
 - **Sort by Name** presents the variable names in alphabetical order.
 - **Sort by Order** presents the variable names in the order in which they appear in the file.
 - b. The TYPE of each variable, such as character or numeric.
 - c. The variable length.
 - d. Whether the variable is an index variable.

- e. Position of the variable within the file (the first variable in the file is at position 0).
 - f. Any formats or informats associated with the variable, such as date formats.
2. When you are finished reviewing the information in this screen, return to the Search/Extract menu screen by using the EXIT pmenu option.
- ! Extracting data to the display table does not make the results “permanent”. Until you save the results as a data set or selection, the extraction results are only temporary and will go away at the end of your StEPS session. You have four options available under the “Process” button located in the display table portion of the screen.
1. **Save in Permanent Data Set:** You will select this option if you wish to make your extracted file a permanent data set to be used in SAS Assist or to be downloaded to your PC. Once you select this option, the following screen will appear, prompting you to select a library and data set name for your new file.

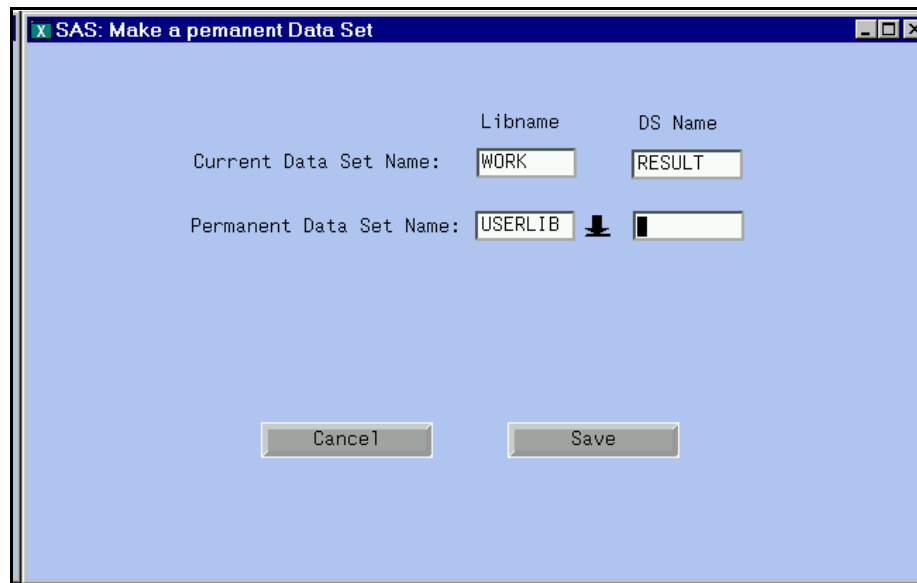


Figure 5.10.4 - Make a permanent Data Set screen

- a. The “Current Data Set Name:” line shows you what StEPS currently calls the new file. Notice that the new data set is located in the temporary WORK directory and has been given the name “RESULT”. Again, since the WORK directory is temporary, the file “RESULT” will disappear once you end your current StEPS session.
- b. You will enter the new library and data set information in the “Permanent Data Set Name:” line.

- i. Notice that StEPS has designated the USERLIB as the potential new location for your extracted file. Unless there is a good reason for your new file to go elsewhere, we recommend you save your file in USERLIB. There is a picklist of available StEPS libraries available if you decide to store your file somewhere else.

Note: You may only store files in existing StEPS libraries. Do not try to make up your own libraries. Use the pick list if you are unsure of the correct library name(s).

- ii. Key the name of your new file into the “DS Name”. We recommend you use a file name that is meaningful and descriptive, as well as easy for you to remember.
 - c. Once you are satisfied with your choices for library and data set names, click the “Save” button. If you wish to discard your choices, click “Cancel”.
2. **Print Data Set:** Use this option to print your new data set. The printed output will be sent automatically to your StEPS default printer (see Chapter 1.1). If your output does not go where you expect, use the WhoAmI screen to check your default printer and change it if necessary.

Caution: Be careful about printing very large data sets. We would recommend printing only after you have significantly reduced the size of your output data set through subsetting. For example, it is usually not a good idea to print out the entire CONTROL file, which can have well over 100,000 records (depending upon the survey).

3. **SAS Insight:** If you select this option, you will be taken directly into SAS Insight and your new data set will be opened in the display table. You may now perform all of the SAS Insight functions on this data set, including saving it to a permanent location.
4. **Create a Selection Set:** By selecting this option, you are saving the list of IDs in your data set as a permanent selection set, which you can then use in the Review and Correction module.

- ! You may obtain summary information on one column of your data by performing a PROC MEANS from the data table. For example, you could use this feature to obtain summary statistics from the EDDATA column of your survey’s item file. Furthermore, using the subsetting features described above, you can obtain summary statistics on just a portion of your items, such as the mean of all the eddata values for ID 1111111111 or all the rpdata values for item CSAL. To access this feature, you need to **left-click** your mouse within a cell for a numeric data column, such as rpdata or eddata from the item file (this feature will not work on character or date fields). Once you have left-clicked in the data field, a pop-up window will appear giving you several options. One of those options will be “Summarize the selected column”. If you single-click on this option, StEPS will perform a PROC MEANS for you and the following output should appear in the SAS output window:

Summary of data set: DATA00.IT1998A1

12:49 Wednesday, August 15, 2001

The MEANS Procedure

Analysis Variable : EDDATA

N	N Miss	Minimum	Maximum	Mean	Sum	Std Error
506870	11744	-9.298877E12	1E13	384728369	1.9500727E14	56286078.68

Figure 5.10.5: Example of output from the PROC MEANS procedure.

1. This output is the same as what you would receive if you ran a PROC MEANS within SAS (outside of StEPS). The information provided in this output is as follows:
 - a. **N** = the number of observations in the data set on which the PROC MEANS was performed. In the example of figure 5.10.5, there are 506870 observations in the NSURV item file for 1998a1. If we had created a subset of the item file before performing the PROC MEANS, this value would have been smaller.
 - b. **N Miss** = the number of missing values in the data column. In our example, there were 11744 missing values within eddata.
 - c. **Minimum** = the smallest data value in the data column, in this case a huge negative number (-9.298877×10^{12})!
 - d. **Maximum** = the largest data value in the data column.
 - e. **Mean** = the mean value in the data column.
 - f. **Sum** = the sum of all the values in the data column.
 - g. **Std Error** = the standard error for the data column.
 2. Remember, the summary values you calculate will be for only one column in your data table. Thus, if you run a PROC MEANS on the eddata column, you will need to run a second PROC MEANS if you want to obtain summary values on the rpdata column.
- ! Another nice feature is the ability to bookmark a particular row in the data table, making it easy to return to that row once you have scrolled on through the data table. To access this feature, left-click any data cell *within the row you are interested in*. A pop-up window will appear with the option “Set Bookmark: row #”. For example, if you are interested in row 34 of the data table, you can left click any data cell within row 34, then click on the “Set Bookmark: Row 34” that appears on the pop-up window. Row 34 is now bookmarked and will remain so until you select another bookmark. If you scroll down to row 24592 and left click on any data cell, you will see two options in the pop-up window: One that says “Set New Bookmark: Row 24592” and another that says “Go To Bookmark: Row 34”. If you select the “Go To” option, you will be returned to Row 34 of the data table. If you select the “Set New” option, Row 24592 will be the new bookmarked row. Row 34 will no longer be bookmarked. Thus, you may only set one bookmark at a time.

P-Menus

P-Menu	Options	Function
HELP	Search By WHERE Help (F1) WhoAmI (F7)	Display HELP information on using the Search/Extract Data by Where Clause screen. Brings up the WhoAmI screen
EXIT	Exit (F3)	Exit to previous screen.